Keyword Weight Propagation for Indexing Structured Web Content

Jong Wook Kim, and K. Selcuk Candan

Comp. Sci. and Eng. Dept
Arizona State University
{jong, candan}@asu.edu
Table

- Motivation
- Approach
- Related Work
- Relative Content of Entries
- Keyword Propagation
  - Keyword Propagation between a Pair of Entries
  - Keyword Propagation across a Complex Structure
- Experiment
- Conclusion and Future Work
Motivation

- Many web sites and portals organize content in a navigation hierarchy
Many web sites and portals organize content in a navigation hierarchy
Motivation

- Many web sites and portals organize content in a navigation hierarchy

- A navigation hierarchy
  - Effective when browsing to find a specific content
  - Semantic relationships between the data contents
    - Generalization/ Specialization
Motivation

- Keyword contents of the intermediate nodes may describe their content in the hierarchy ambiguously.

The Yahoo CS hierarchy
Motivation

- In a navigational hierarchy, keyword searches are usually directed
  - to the root of the hierarchy, or
    - Undesirable topic drift

- to the leaves
  - May not be enough to satisfy the query

- It is important for individual nodes to be properly indexed
Table

- Motivation
- Approach
- Related Work
- Relative Content of Entries
- Keyword Propagation
  - Keyword Propagation between a Pair of Entries
  - Keyword Propagation across a Complex Structure
- Experiment
- Conclusion and Future Work
Approach

- **Keyword and keyword weight propagation**
  - Enrich the individual nodes with the contents of the neighboring nodes

- **How to decide what to propagate and how much?**
  - The original semantic structure should be preserved
    - Generalization/ Specialization

- **Challenge**
  - How to represent the semantic structure (i.e., generalization/specialization) between nodes?
  - How to determine the degree of keyword inheritance?
Approach

Contributions of the Paper

- Develop a method for discovering and quantifying the generalization/specialization relationship between entries in a navigation hierarchy

- Develop a keyword propagation algorithm using this relationship
Table

- Motivation
- Approach
  - Related Work
- Relative Content of Entries
- Keyword Propagation
  - Keyword Propagation between a Pair of Entries
  - Keyword Propagation across a Complex Structure
- Experiment
- Conclusion and Future Work
Related Work

- **Score and Keyword Frequency Propagation**
  - Propagate the relevance score [Shakery, and Zhai, TREC’03]
  - Propagate the term frequency value [Savoy et al. JASIS’97]
  - [Song et al. TREC’04]
  - Propagate the relevance score and the term frequency value [Qin et al. SIGIR’05]
Table

- Motivation
- Related Work
- Approach
  - Relative Content of Entries
- Keyword Propagation
  - Keyword Propagation between a Pair of Entries
  - Keyword Propagation across a Complex Structure
- Experiment
- Conclusion and Future Work
Relative Content of Entries

- In a navigation hierarchy,
  - A specialized entry corresponds to more constrained concept
    - As one moves down in a hierarchy, the nodes get more specialized
  - A general entry is less constrained
    - As one moves up in a hierarchy, the nodes get more generalized.
Intuition

Given two entries, A and B (A is an ancestor of B),

- Assume
  - A has three keyword (k1, k2, k3), and
  - B has two keyword (k2, k3)

- “Entry A is more general than B” → A being less constrained than B by keywords

- If B is interpreted as k2 ∨ k3, then A should be interpreted as k1 ∨ k2 ∨ k3
  - Less constrained than k2 ∨ k3

- Interpreted as the disjunction of keywords
In extended boolean model [Salton 83],

- **OR-ness**
  - An entry further away from $O$ better matches the $k1 \lor k2$
  - Measured as a distance from $O$

$$O = \neg (k1 \lor k2)$$
Relative Content of Entries

Given two entries, A and B (A is an ancestor of B),

Assume

- A has three keyword (k1, k2, k3), and
- B has two keyword (k2, k3)

How much entry A and B represent a disjunct?

\[ |\overrightarrow{A} - \overrightarrow{O}| >= |\overrightarrow{A}|, \quad |\overrightarrow{B} - \overrightarrow{O}| >= |\overrightarrow{B}| \]

If A is more general than B, then

\[ |\overrightarrow{A} - \overrightarrow{O}| > |\overrightarrow{B} - \overrightarrow{O}| \]
Relative Content of Entries

- **Visual representation of the keyword contents**

- **Relative Content**

\[
R_{AB} = \frac{|\overrightarrow{A}|}{|\overrightarrow{B_C}|} = \frac{|\overrightarrow{A_U} + \overrightarrow{A_C}|}{|\overrightarrow{B_C}|}
\]

Measure whether the additional keywords (A_u) make A more general or less general than B_C
Table

- Motivation
- Approach
- Related Work
- Relative Content of Entries
- Keyword Propagation
  - Keyword Propagation between a Pair of Entries
  - Keyword Propagation across a Complex Structure
- Experiment
- Conclusion and Future Work
Keyword Propagation between a pair of entries

- The purpose of keyword propagation
  - Enrich the entries in a navigational hierarchy
  - The original semantic properties (i.e., relative generality) should be preserved

- Propagation Degree, $\alpha$
  - Govern how much keyword weights two neighboring entries should exchange
Keyword Propagation between a pair of entries

- **Propagation Degree, \( \alpha \)**
  - Given two entries, \( A \) and \( B \),
    - \( a_i : \) weight associated with keywords \( k_i \in K_A \)
    - \( b_i : \) weight associated with keywords \( k_i \in K_B \)

- **\( A' \) and \( B' \)**
  - Enriched entries after keyword propagation

- **For all \( k_i \in K_A' \)**
  - \( \star \) If \( k_i \in (K_A - K_B) \), then \( a_i' = a_i \)
  - \( \star \) If \( k_i \in (K_A \cap K_B) \), then \( a_i' = a_i + \alpha b_i \)
  - \( \star \) If \( k_i \in (K_B - K_A) \), then \( a_i' = \alpha b_i \)

- **For all \( k_i \in K_B' \)**
  - \( \star \) If \( k_i \in (K_A - K_B) \), then \( b_i' = a a_i \)
  - \( \star \) If \( k_i \in (K_A \cap K_B) \), then \( b_i' = b_i + a a_i \)
  - \( \star \) If \( k_i \in (K_B - K_A) \), then \( b_i' = b_i \)
Keyword Propagation between a pair of entries

- **Propagation Degree, $\alpha$**
  - $A'$ and $B'$ are located in a common keyword space
    \[ K_C = K_{A'} = K_{B'} = K_A \cup K_B \]
  - After keyword propagation, relative content should be preserved

\[ R_{A'B'} = R_{AB} \]

\[ R_{A'B'} = \frac{|\vec{A}|}{|\vec{B}_C|} = \frac{|\vec{A'}|}{|\vec{B'}|} = R_{AB} \]
Table

- Motivation
- Approach
- Related Work
- Relative Content of Entries
- Keyword Propagation
  - Keyword Propagation between a Pair of Entries
    - Keyword Propagation across a Complex Structure
- Experiment
- Conclusion and Future Work
Keyword Propagation across a Complex Structure

- Let $H(N,E)$ be a navigation hierarchy,
  - $N$: the set of nodes
  - $E$: the set of edges

- Propagation Adjacency Matrix, $M$
  - If there is an edge $e_{ij} \in E$, then both $(i,j)$ and $(j,i)$ of $M$ are equals to $\alpha_{ij}$ (the pairwise propagation degree)
  - Otherwise, both $(i,j)$ and $(j,i)$ of $M$ are equal to 0.

![Diagram of a navigation hierarchy with nodes n1, n2, and n3 connected by edges $\alpha_{12}$ and $\alpha_{23}$, and the propagation adjacency matrix.

$\begin{pmatrix}
0 & \alpha_{12} & 0 \\
\alpha_{12} & 0 & \alpha_{23} \\
0 & \alpha_{23} & 0
\end{pmatrix}$

WebKDD 2006 Workshop on Knowledge Discovery on the Web, Aug. 20, 2006,
Philadelphia, PA, USA
Keyword Propagation across a Complex Structure

Keypropagation Process

- Given a hierarchy, $H(N,E)$
  - $T$ : Term-node matrix
  - $M$ : Propagation Adjacency matrix

Term Propagation Matrix

$$P = T M$$

$$P = \begin{pmatrix} 0 & a_{12} K1 & 0 \\ a_{12} K2 & a_{12} K2 & a_{23} K2 \\ a_{12} K3 & a_{23} K3 & a_{23} K3 \end{pmatrix} = \begin{pmatrix} K1 & 0 & 0 \\ K2 & K2 & 0 \\ 0 & K3 & K3 \end{pmatrix} \begin{pmatrix} 0 & a_{12} & 0 \\ a_{12} & 0 & a_{23} \\ 0 & a_{23} & 0 \end{pmatrix}$$

Inherited from its neighbors in $M$
After keyword propagation

\[ T' = T + P = T + TM = T(I + M) = TM \]

- New enriched term-node matrix
- Propagation Adjacency matrix
- All diagonal values are 1 and all non-diagonal entries are same with M
Keyword Propagation Process

(a) Original content

(b) After the first keyword propagation

(c) After the second keyword propagation
Keyword Propagation Process

(a) Original content

(b) After the first keyword propagation

(c) After the second keyword propagation

\[ T_{\text{final}} = TM_1M_2 \ldots M_d \]

Propagation adjacency matrix computed for the \(d^{th}\) iteration

\(d\) is the greatest number of edges between any nodes

WebKDD 2006 Workshop on Knowledge Discovery on the Web, Aug. 20, 2006, Philadelphia, PA, USA
Table

- Motivation
- Approach
- Related Work
- Relative Content of Entries
- Keyword Propagation
  - Keyword Propagation between a Pair of Entries
  - Keyword Propagation across a Complex Structure
- Experiment
- Conclusion and Future Work
Experiment

- **Experiment Setup**
  - **Data**
    - Yahoo Hierarchy
    - Computer Science, Mathematics, and Movie directory
  - **Ground truth and Query**
    - 10 sample keyword queries
    - User study (8 users)

<table>
<thead>
<tr>
<th>r</th>
<th>Relaxed</th>
<th>Differentiated</th>
<th>Strict</th>
</tr>
</thead>
<tbody>
<tr>
<td>irrelevant</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>partially relevant</td>
<td>1</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>fully relevant</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Experiment

- **Experiment Setup**
  - **Query processing**
    - N (No Keyword Propagation)
    - KP (Keyword Propagation)
    - D_t and D_n
      - No Keyword Propagation, but context extracted from the whole tree or neighbor
    - KP+ D_t and KP31+D_n
      - Keyword Propagation, and context extracted from the whole tree or neighbor

- **Evaluation measure**
  - P@10
  - MRR (Mean reciprocal rank of the first relevant document)
  - Paired t-Test
## Keyword Propagation/ No Propagation

<table>
<thead>
<tr>
<th></th>
<th>( N )</th>
<th>( KP )</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relaxed</td>
<td>0.670</td>
<td>0.753</td>
<td>12.27%</td>
</tr>
<tr>
<td>Differentiated</td>
<td>0.542</td>
<td>0.612</td>
<td>12.60%</td>
</tr>
<tr>
<td>Strict</td>
<td>0.415</td>
<td>0.469</td>
<td>13.10%</td>
</tr>
</tbody>
</table>

**P@10**

<table>
<thead>
<tr>
<th></th>
<th>( N )</th>
<th>( KP )</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relaxed</td>
<td>0.869</td>
<td>0.930</td>
<td>6.97%</td>
</tr>
<tr>
<td>Differentiated</td>
<td>0.869</td>
<td>0.930</td>
<td>6.97%</td>
</tr>
<tr>
<td>Strict</td>
<td>0.644</td>
<td>0.730</td>
<td>13.20%</td>
</tr>
</tbody>
</table>

## Average MRR
Keyword Propagation/ No Propagation

<table>
<thead>
<tr>
<th>p-values for $KP \text{ vs. } N$</th>
<th>Relaxed</th>
<th>Differentiated</th>
<th>Strict</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.029</td>
<td>0.031</td>
<td>0.047</td>
</tr>
</tbody>
</table>

P-values for the t-Test
Keyword Propagation/ No Propagation

(a) Relaxed Precision vs. Ranking

(b) Differentiated Precision vs. Ranking

(c) Strict Precision vs. Ranking
## Keyword Propagation/Alternative Context Extraction

<table>
<thead>
<tr>
<th>Differentiated: P@10</th>
<th>( \beta/\gamma )</th>
<th>1/0</th>
<th>0.8/0.2</th>
<th>0.6/0.4</th>
<th>0.4/0.6</th>
<th>0.2/0.8</th>
<th>0/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>0.542</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( D_t )</td>
<td>-</td>
<td>0.539</td>
<td>0.545</td>
<td>0.579</td>
<td>0.558</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>( D_n )</td>
<td>-</td>
<td>0.532</td>
<td>0.542</td>
<td>0.547</td>
<td>0.564</td>
<td>0.572</td>
<td></td>
</tr>
<tr>
<td>KP</td>
<td>0.612</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KP+( D_t )</td>
<td>-</td>
<td>0.606</td>
<td>0.607</td>
<td>0.607</td>
<td>0.597</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>KP+( D_n )</td>
<td>-</td>
<td>0.611</td>
<td>0.612</td>
<td>0.596</td>
<td>0.584</td>
<td>0.572</td>
<td></td>
</tr>
</tbody>
</table>

### Differentiated: t-Test

<table>
<thead>
<tr>
<th>Differentiated: t-Test</th>
<th>( \beta/\gamma )</th>
<th>1/0</th>
<th>0.8/0.2</th>
<th>0.6/0.4</th>
<th>0.4/0.6</th>
<th>0.2/0.8</th>
<th>0/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_t ) vs. ( N )</td>
<td>-</td>
<td>worse</td>
<td>55.1%</td>
<td>84.4%</td>
<td>63.5%</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>( D_n ) vs. ( N )</td>
<td>-</td>
<td>worse</td>
<td>54.0%</td>
<td>65.2%</td>
<td>81.1%</td>
<td>90.5%</td>
<td></td>
</tr>
<tr>
<td>KP vs. ( N )</td>
<td>96.9%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>KP+( D_t ) vs ( N )</td>
<td>-</td>
<td>96.2%</td>
<td>95.7%</td>
<td>95.7%</td>
<td>90.0%</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>KP+( D_n ) vs ( N )</td>
<td>-</td>
<td>96.7%</td>
<td>96.8%</td>
<td>91.8%</td>
<td>86.5%</td>
<td>90.5%</td>
<td></td>
</tr>
</tbody>
</table>

### Differentiated: t-Test relative

No Keyword Propagation
Effect of the Structural Distance
Statistical Validation of the Ground Truth

- **ANOVA test**
  - A statistical test to observe the agreement between the assessors

- We identified two users whose judgments were significantly different from the other 6 users

- When excluding these two users, the user judgments were in agreement
# Statistical Validation of the Ground Truth

## Differentiated: P@10

<table>
<thead>
<tr>
<th>(\beta/\gamma)</th>
<th>1/0</th>
<th>0.8/0.2</th>
<th>0.6/0.4</th>
<th>0.4/0.6</th>
<th>0.2/0.8</th>
<th>0/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N)</td>
<td>0.538</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(D_t)</td>
<td>-</td>
<td>0.547</td>
<td>0.556</td>
<td>0.594</td>
<td>0.571</td>
<td>NA</td>
</tr>
<tr>
<td>(D_n)</td>
<td>-</td>
<td>0.525</td>
<td>0.537</td>
<td>0.544</td>
<td>0.565</td>
<td>0.573</td>
</tr>
<tr>
<td>KP</td>
<td>0.628</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KP+(D_t)</td>
<td>-</td>
<td>0.625</td>
<td>0.624</td>
<td>0.624</td>
<td>0.608</td>
<td>NA</td>
</tr>
<tr>
<td>KP+(D_n)</td>
<td>-</td>
<td>0.625</td>
<td>0.625</td>
<td>0.614</td>
<td>0.601</td>
<td>0.573</td>
</tr>
</tbody>
</table>

## Differentiated: t-Test

<table>
<thead>
<tr>
<th>(\beta/\gamma)</th>
<th>1/0</th>
<th>0.8/0.2</th>
<th>0.6/0.4</th>
<th>0.4/0.6</th>
<th>0.2/0.8</th>
<th>0/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D_t) vs. (N)</td>
<td>-</td>
<td>62.0%</td>
<td>71.3%</td>
<td>80.4%</td>
<td>72.1%</td>
<td>NA</td>
</tr>
<tr>
<td>(D_n) vs. (N)</td>
<td>-</td>
<td>worse</td>
<td>worse</td>
<td>69.9%</td>
<td>74.9%</td>
<td>91.8%</td>
</tr>
<tr>
<td>KP vs. (N)</td>
<td>97.3%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KP+(D_t) vs (N)</td>
<td>-</td>
<td>96.4%</td>
<td>96.6%</td>
<td>96.6%</td>
<td>90.5%</td>
<td>NA</td>
</tr>
<tr>
<td>KP+(D_n) vs (N)</td>
<td>-</td>
<td>96.4%</td>
<td>96.4%</td>
<td>93.8%</td>
<td>90.5%</td>
<td>91.8%</td>
</tr>
</tbody>
</table>

### Differentiated: t-Test relative

**No Keyword Propagation**

---

WebKDD 2006 Workshop on Knowledge Discovery on the Web, Aug. 20, 2006, Philadelphia, PA, USA
Table

- Motivation
- Approach
- Related Work
- Relative Content of Entries
- Keyword Propagation
  - Keyword Propagation between a Pair of Entries
  - Keyword Propagation across a Complex Structure
- Experiment
  - Conclusion and Future Work
Conclusion and Future Work

- **Conclusion**
  - Present a technique to identify a semantic relationship
  - Introduce a relative content preserving keyword propagation technique

- **Future Work**
  - Incorporate of other types of semantic cues
    - Structured-based method
    - Information-based method
Question